THE RETRIEVAL PROBLEM FOR HEALTH POLICY AND PUBLIC HEALTH: KNOWLEDGE BASES AND SEARCH ENGINES

CLIFFORD LYNCH, PHD

INTRODUCTION: INFORMATION NEEDS AND RESOURCES

Other papers presented at this conference and reported in this issue of the *Journal of Urban Health* examined in depth the specific information needs and resources of health policy and public health professionals. This paper explores and assesses computer-based searching tools and systems that are available to these communities. Public health professionals will need to know how to use the growing array of information resources appearing on the Internet, some of which may specifically address their professions, but many of which were created primarily to serve other purposes. The linkage of public health and health policy to geographic communities guarantees a continuing requirement for access to diverse resources that will vary from one professional to another. This paper highlights a few developments in other disciplines, in which networked information resources and tools are making a significant difference in problem solving and general research, as a way of illustrating what may be possible for the public health communities through the application of these same technologies. Many of the most prominent examples (which we do not discuss, however) are in the biomedical and life sciences.

One goal of this paper is to provide readers with a realistic understanding of what they can expect—today and in the near future—from the Web and the Internet. Readers will learn how these tools can enhance, supplement, or even replace existing resources and practices and thus understand the extent to which investments in both creating and learning about how to use information technol-

Dr. Lynch is Executive Director, Coalition for Networked Information, 21 Dupont Circle, NW, Suite 800, Washington, DC 20036. (E-mail: cliff@cni.org)

ogy, network connectivity, and electronic information resources should be a priority for the public health and health policy communities. I also discuss some of the limitations of these tools. Even though the Internet and the Web have facilitated enormous advances in information access, the state of the art is still in many ways surprisingly primitive and will likely improve only slowly and incrementally. The problems are not only technical; there are also social, economic, business, and legal aspects, all of which are compounded by the sheer scale of the Internet and the complexities this represents as a deployment environment.

I am not a public health or health policy professional. My training and experience are in the computer and information sciences and in library automation. I rely heavily on the work of my fellow presenters at this conference to characterize the information discovery problems of the public health community and the sorts of information sources that are needed to meet these needs. However, I take all responsibility for any incorrect assumptions or misunderstandings stated in this paper, and I ask that readers assess my comments on specific public health and health policy information resources with greater skepticism than they might my comments about information technology.

In general, I think that the information-seeking approach of public health and health policy professionals can be characterized in a few important ways:

- It is generally focused on problem solving rather than on answering open-ended academic research questions.
- It involves a synthesis of information from many sources and many disciplines; often it involves giving a new purpose to information that was originally created and organized to serve other communities.
- There is often considerable time pressure for information; in addition, the health professionals are often extremely busy and place a high premium on being able to obtain accurate, comprehensive, and timely information quickly.
- The base of information needed goes far beyond traditional published literature and encompasses a wide range of fugitive information and information created as a by-product of organizational and institutional operations. It also goes beyond textual information to include, for example, statistical and geospatial data sets. Because of the range of information sources and types involved, assessments of quality, accuracy, and timeliness by health professionals is an integral part of the information discovery and synthesis process.
- Budgets to acquire information are limited; there is not a large commercial marketplace in developing information resources targeted specifically to

the needs of these communities, although there are some critical resources available from the public sector. Put another way, it seems unlikely that the private sector will engineer a comprehensive system to respond to the needs of health policy and public health professionals in the near future (in contrast to attorneys, for example). Instead, meeting these needs will be a matter of employing a range of existing and developing systems and resources, many of which are being targeted for other audiences.

- Public health and health policy professionals do not always have easy access to the resources of a major research library; thus, the promise of network-based access to comprehensive information resources may alter substantially what is practical for some professionals.
- For many professionals, there is both a common disciplinary and a unique regional or local component to their information needs because of the strong local/regional basis for much of the health policy and public health system.

Given this background, it is clear that the networked information environment offers great promise. It certainly contains an enormous diversity of information that goes far beyond the traditional published literature, although it does not necessarily incorporate that traditional published literature in a systematic way. Access is available without the constraints of geographic proximity to a library. The Internet has totally revolutionized the economics of access to certain types of information, such as data generated by the federal government. Patents, Securities and Exchange Commission financial filings, Federal Communications Commission rulings, and similar materials once were available only at high prices to a relatively elite user community (or through a library) and are now accessible to the general public easily and for free. Yet, it should be clear already that the Internet does not represent a comprehensive, engineered solution to the information needs of public health and health policy so much as it offers certain opportunities to resolve various components of these needs on a somewhat patchwork basis.

This paper first provides a broad framework for thinking about the Internet and the Web as comprising an information access environment. It then discusses a number of systemic information access problems in the networked environment and offers some comments on specific classes of information that are important to the public health and health policy communities. The paper closes with a brief consideration of other roles of the networked information environment in facilitating communication and collaboration rather than simply providing

information access and how this may change our thinking about information discovery.

THE REALITIES OF ELECTRONIC INFORMATION ACCESS VIA THE INTERNET

From the perspective of information access, we can think of the Web and the Internet in two distinct ways. The distinction is subtle, but of crucial importance. It is not absolute: there are two points of view, and many sites can be viewed from both perspectives.

First, the World Wide Web and the Internet comprise a vast collection of digital documents that various individuals and organizations have made available. If one knows about a specific document (e.g., if one knows the uniform resource locator, or URL, of an object), one can examine it through a Web browser such Netscape Navigator, NCSA Mosaic, Opera, or Microsoft Internet Explorer. If one does not know the specific document desired, one can search some parts of this collection of documents through Web indexing services such as AltaVista, Lycos, or HotBot, or one can navigate through Web cataloging services such as Yahoo! to find sites that deal with specific subjects. I return to the specific mechanics of these organizing systems below. Note that I use the term *document* in a very general way to include not only text, but also data sets and digital multimedia.

The other way of thinking about the Web and the Internet is as an access mechanism and user interface to a wide range of interactive information resources, such as databases or journals in electronic form. If one knows a specific site that houses a database—for example, MEDLINE—one can interact with that site by filling out forms in the Web browser to construct and submit queries and later receive results formatted as Web pages. Different sites offer different searching capabilities, and the content hosted by different sites is organized in a wide variety of ways that are specific to their content. In some cases, older interactive services are designed to work with character-based terminals, requiring the use of a terminal emulator and the Telnet protocol, rather than Web forms, to interact with them. The key point is that one uses the Internet to reach a resource and the Web browser to structure one's interaction with it. The content managed by the resource may not be part of the vast collection of digital documents on the Web and may not be visible to the Web indexing services. One gradually learns what sites offer what resources, primarily through publicity that the site or third parties have issued or through directories and catalogs of information resources.

At the most primitive level, any Web site can be viewed as an interactive

information resource, even if it is only a set of linked Web documents, perhaps supplemented by a local search engine that covers the site. To the extent that a site is more than this, however, it is increasingly useful to view it as an interactive information resource because the content it hosts is not part of the document-collection Web.

One might think of these two ways of thinking about the Web as two aspects of the Web or even as two separate Webs of information that coexist and share common usage conventions established by the idioms of hypertext and the Web browser.

The Web as a Document Collection

It is inaccurate and dangerous to think of the Web as the world's library. Many things that one would expect to find in a library are not on the Web, certainly not the document-collection Web. A tremendous amount of material that is on the Web would never be part of a traditional library: preprints, wedding invitations, scrapbooks and photo albums, meeting minutes and announcements, and press releases. Anyone can place material on the Web; there is no vetting process, little accountability, and no guarantee that information will be maintained once it is placed there. We still seem to expect that Web sites are maintained, even while we are comfortable with the idea that a print publication represents a snapshot at a specific point in time. Perhaps the best way to think about the Web as a document collection is as a very generous random sampling of the output of the world's digital printing presses (as distinct from publishers), with the sample biased by omitting most material of well-established high commercial value that is owned by the private sector.

Finding materials of interest within the document-collection Web is problematic at best. Some services, such as Yahoo!, identify content (mostly at the site level) using practices similar to the selection and cataloging that goes on in libraries—though the content emphasis is, not surprisingly, slanted toward shopping, entertainment, personal business, and popular culture rather than science and scholarship. These services pay people to trawl the Web continually, looking for good sites to catalog and describe according to a classification scheme. There are also an enormous number of more-specialized pages that serve as bibliographies or catalogs for materials on specific subjects; but even finding these bibliographic or catalog pages can be difficult and haphazard. Coverage of content (of either Web resources or Web documents) on the Web through intellectual description is limited for a number of reasons. Human cataloging is costly, and (particularly in the advertising-support funding model used by Yahoo!)

economics dictate that enough people will be available to examine only a small fraction of the material on the Web. The bias toward cataloging sites—the interactive information resource rather than the document view of the Web—is understandable because there are so many fewer sites to visit than there are documents, and the sites are more stable than the document URLs.

Some of the abstracting and indexing services that have traditionally covered specific scientific literature are now starting to encompass materials on the Web as well. It remains to be seen, however, how timely and comprehensive their coverage will be. It is much easier and cheaper to make a one-time determination that a specific scholarly journal has an appropriate scope and level of quality to merit coverage in an abstracting and indexing database than it is to evaluate a dynamic Web site continually for ongoing inclusion in an abstracting and indexing service.

The other major approach to providing access to the document-collection Web is through Web indexers. These systems, such as AltaVista, Lycos, Infoseek, Excite, and HotBot constantly run computer programs that visit sites on the Web, looking for documents and extracting key words, which then are placed in index databases. While these services are tremendously valuable, they suffer from very real limitations. First, it is difficult to know what is included and how quickly new materials are indexed. The programs take a long time to cycle through the Web and to discover new documents (or to discover that old documents have disappeared). They all have proprietary means for finding sites to index; the best guess is that currently they collectively cover somewhat less than half of the material on the Web. Many of the indexing services only sample sites; they do not necessarily index every page on the site (and often they are rather secretive about their policies in this area). They rely on the fact that the site is willing to have foreign programs index it; most publishers who license access to materials on the Web will not permit these programs to index their materials because they want to protect their proprietary content. The programs cannot "see inside" databases or other retrieval systems; thus, they are limited to the document-collection Web.

An example may clarify this structure. The Securities and Exchange EDGAR site has many millions of documents that represent corporate financial filings in a database; it allows users to search the database via a Web form. However, a Web indexer visiting the EDGAR site will only find a few pages of introductory materials and help screens to index; everything else is in the database. Therefore, if one uses "SEC" as a search term, one will find the EDGAR site, but if one is looking for "IBM," one will likely not find the EDGAR site, despite the fact that

it holds numerous filings about IBM. Similarly, Elsevier Science Publishers is building a site with over 1,000 journals; only a few welcome screens, and none of the contents of these journals, will be indexed in the Web indexing services. Many huge interactive information resources are part of this invisible Web, however, such as Lexis/Nexis, the American Chemical Society, and patent and molecular biology databases.

The other major limitation of Web indexing systems today is that they work purely by computer manipulation of unstructured text. They do not use a systematic vocabulary to describe content; they rely only on the words that appear in a document. They cannot describe a document that uses one terminology in another or describe in English what a document in Spanish is about. Searching using specialized criteria and vocabulary (for example, chemical components and structures) is basically impossible. The Web indexers do not know how to identify and index this type of content. They cannot describe images, video clips, simulation models, or data sets other than to the extent of extracting words from textual pages that might be linked to the images or data sets. They use statistical measures to guess the primary topic of a document. While this process does allow one to find *any* indexed page that mentions a particular person or place (if one has the patience), the Web indexing systems often do not guide users rapidly to the key documents that are connected most closely to the topic of a given search.

Because they are designed to work with unstructured documents, these indexers can support only very limited types of searching. For example, one cannot differentiate documents *authored by* a person from documents *about* a person using current Web indexers because the indexers have no way of telling. All they know is that the name in question appears in the document. This may begin to change over the next few years as initiatives such as the Dublin Core Metadata Program and the World Wide Web Consortium's Resource Description Framework begin to deploy on the Web, though it remains to be seen how much structure most authors are willing to apply to their documents.

Finally, it is important to recognize that navigating the document-collection Web is about *finding documents*. Interactive information resources can be designed to be much more flexible and to answer a wide range of questions. One can design a database that can tell the interested seeker the temperature in New York City on January 1, 1997, or the boiling point of water, or the number of people in a specific county that had chicken pox last year. The document-collection Web cannot answer these questions directly, but it can help one to find documents from which one may be able to extract the answers. The document-collection

Web is not a global distributed knowledge base except in the most abstract and indirect way and when mediated by human intelligence.

The Net as Access to Information Resources

If we take the view of the Internet as a collection of independent interactive information resources rather than a collection of documents, then we are faced with two key issues: finding the right resources (sites) from which we can search or otherwise interact and using the resources successfully.

As suggested above, the only feasible way to organize interactive information resources at present is through human intellectual analysis. We have little production-ready technology today that effectively lets a computer program analyze an arbitrary information resource the way that it can analyze and index a document by key words. Further, some of these resources are highly proprietary, charging hundreds of dollars per hour for access. Their owners will not permit them to be indexed by external programs even if this were possible.

Finding the right resources to search is very different from finding documents themselves. It requires a much greater understanding of how information is viewed and organized. Let us return to the case of EDGAR as an example. One would see EDGAR listed as a financial data resource; it would not be listed under each company that files with it. The information seeker would have to know what kinds of companies file data with EDGAR and whether those data were relevant when researching a specific company. To assess resources fully, one needs to know not only the types of material that they contain, but also whether it is authoritative, current, and comprehensive. Coverage and inclusion policies for information resources are complex and change over time, making such assessment difficult.

The second issue is the actual use of the resource. Many of these resources offer very sophisticated searching facilities, but they come with a learning curve. Each resource tends to be unique. Different databases may employ specialized controlled vocabularies, which take time to learn. Unlike the document Web, for which the majority of the available content is free for the reading, many interactive resources charge fees, using complex algorithms, and require complex use agreements.

SYSTEMIC ACCESS ISSUES

After defining a framework for understanding Internet-based information access, it may be valuable to see how that framework corresponds to the broader systemic information access problem.

At one time, most information was in printed form. Abstracting and indexing

databases provided access to a selected subset of this information. To the extent that they organized the right body of information in ways that were useful to people looking for information, access was relatively straightforward. The problems that public health and health policy professionals faced were two-fold: they had to employ multiple abstracting and indexing databases because of the scope of information they needed (with all of the overlap and terminology problems that implied) and, because much of the key information that they needed was not part of traditional published literature, it was not organized by any abstracting and indexing database. They had to deal with this fugitive literature in a relatively *ad hoc* fashion.

Applying computer technology to search these abstracting and indexing databases simply made searching more powerful and convenient. The retrospective literature was covered only by printed abstracting and indexing tools, but not by the new and much more convenient computerized databases. As a result, recent literature was much more visible and accessible than the historical literature.

Now, as an increasing amount of primary content (and not just the abstracting and indexing databases that traditionally have described it) becomes electronic in the networked information environment, we have the following situation:

- Abstracting and indexing services continue to organize the traditional literature; some of it is electronic and some of it is print. Some abstracting and indexing services also are starting to consider organizing content that goes beyond the traditional literature. Nothing much has changed, except that sometimes one can get the literature being described electronically rather than in print, and that as abstracting and indexing databases proliferate, it is harder to determine which ones are appropriate to use for a given search. Most abstracting and indexing databases now can be conveniently reached via the Internet and searched using Web-based interfaces.
- A considerable amount of the nontraditional literature that is of interest to the public health and health policy communities now is available in electronic form. Unfortunately, it is hard to know what is available electronically, what is still only in print, and what can be found in both forms; this situation is changing day to day.
- Some of the nontraditional literature and other information can be located through Web search engines. The quality of access is very different from what we have come to expect for the published literature organized via abstracting and indexing databases. In a few ways, it is better—for example,

the ability to find every document that mentions a name—but in most ways, it is far less precise and predictable.

- Other components of the fugitive information base important to public health and health policy now are accessible through new interactive information resources that can be reached through the Internet. However, one must be able to find the appropriate resources (which can be difficult), learn how to use them, and be prepared to pay high rates for access to some resources that are commercial in nature.
- The availability of the Web is encouraging new genres of content, such as Web sites, frequently-asked-question lists, and simulation models, and larger quantities of existing forms of content, such as press releases, which are important information sources for public health and health policy. The much more distributed and democratic nature of "publishing" in the Web environment also means that critical assessment of quality is much more important for some of these materials.

This electronic world is a much more fragmented place; it is a world in which an information seeker is likely to be less confident about finding all of the relevant materials.

There are two key questions that the public health and health policy communities need to explore:

1. What is the mix of use of the document-collection Web and the Internet as a means of accessing complex interactive information resources? Can some specific parts of the document-collection Web be brought under the control and better organization offered by custom-designed interactive information resources developed for the public health and health policy communities?
2. To the extent that interactive information resources on the Internet rather than Web documents are used, what specialized catalogs or other guides might be constructed to aid members of the community in locating the appropriate resources for various purposes? Apparently, some efforts already are under way in this area.

The answers to these questions will help identify options for community-wide action to leverage investment.

COMMENTS ON SPECIFIC CLASSES OF INFORMATION

There are a few specific classes of information that deserve special comment because of either their unique characteristics or their particular importance to the

needs of public health and health policy professionals: government information, geospatial information, and "social" or "community" information.

Government Information

A tremendous amount of information generated by the federal government is available on the Internet, including both legislative information and information from the executive branch and federal agencies. Many agencies now are using networked information as a centerpiece of their information-distribution strategies. Some work also has been done on GILS—the Government Information Locator System—which provides a high-level tool for identifying information resources held by the federal government, although a recent evaluation of this system has suggested there are still some substantial problems to be addressed.

While there are a few surprising gaps—for example, the majority of the body of federal case law is available only from costly private services—the bounty of federal government information available on the Internet is largely a result of a set of information policies that dictate that federal information be unprotected by copyright because it is in the public interest to make it widely available.

This is most emphatically not the case for most state, regional, and local information. Information policy and practice at this level varies tremendously from locale to locale; some state and local governments regard access to this information as an increasingly important source of revenue. The commitment and expertise necessary to move to electronic information sources is also uneven, so it is very difficult to know what to expect in terms of availability and costs for state and local "public" information resources.

Documentation of the processes of government at all levels—meeting agendas and minutes, hearings, regulations, and the like—is also is starting to appear on the Web.

Geospatial Information

Substantial resources now are being invested in the construction of what is termed the "National Spatial Data Infrastructure" and much geospatial information is becoming available on the Web. The majority of this, because of its nontextual information, is stored within network-accessible interactive information resources. While the amount of information available is substantial, it is difficult to integrate with other information from other sources without a good deal of expertise and specialized Geographic Information System software.

Statistical Information

The majority of the available statistical information is housed in specialized information resources; the body of statistical information on the Web is growing rapidly. Knowing where to search is essential. There are no tools of which I am aware to help people determine what statistical data sets might be available. At best, there are fairly general descriptions of the statistical data sets available, but to really understand them, one needs to research data set groups by examining codebooks.

"Social" or "Community" Information

One of the most powerful capabilities of the Internet is its ability to host innumerable communities of interest at relatively low cost and to facilitate communication and information sharing among members of those communities. There are now active networked "disease communities" composed of patients, concerned family members, advocacy and support groups, and medical professionals; these communities serve as forums for those concerned with the management of diseases both commonplace and rare. These groups share information, rumor, data on clinical trials, alternative medicine information, and coping strategies. Part of the activities of these groups is reflected in Web sites that can be found in the traditional ways. These communities also may make extensive use of mailing lists or newsgroups, and they are as much concerned with communication among their members as they are with building a lasting community reference resource. In the digital environment, however, all communication has the potential to create a new information resource, which records that communication. The archives of these mailing lists or newsgroups may be important information sources for the public health and health policy communities. There are specialized tools (such as the Deja News indexing service) for searching newsgroup archives that are separate from the more traditional Web search mechanisms. Mailing list archives may be placed on the Web (and thus indexed through the usual Web indexing mechanisms), but often are not. Some special effort may be worthwhile to obtain access to these archives.

CONCLUSION: NEW ASPECTS OF THE NETWORKED INFORMATION ENVIRONMENT

It is important for the public health and health policy communities to recognize that the networked information environment should not be viewed merely as a new way to search the traditional literature and to gain improved access to the fugitive materials and other information sources that they need. Something much more complex and significant is happening, particularly as we look not only at

current technology, but also at developments that are likely to emerge in the next few years.

Personalization and Integration

At present, most information seeking on the Internet is integrated poorly with other activities. The model is a decontextualized, user-initiated searching process as a discrete action rather than a continual uncovering of relevant information from the Internet in a context of past actions, interests, and work processes. More-sophisticated tools that retain context, learn more about users, and use that knowledge to mediate users' information seeking on an ongoing basis are likely to appear. These tools will facilitate better the continued interaction between a limited set of key information resources and the working professional that is more typical of workplace information use than today's user-initiated search model. We will see systems for document and work-flow management, calendaring, and other work-activity support become better integrated with information management and seeking systems.

Communities, Communication, and Collaboration

The Internet is not just a way of publishing and retrieving information. It is a set of tools for communicating among people, for building communities, and for collaborating among groups. Technically simple tools like listservers and e-mail reflectors already have had a very powerful social impact. The public health and health policy communities interact with, serve, and are accountable to a very wide range of communities. The networked environment provides opportunities not just for richer interaction within the professions, but also for richer interaction with these external groups.

Simple e-mail communication will be supplemented within the next decade by very sophisticated multimedia-based collaborative authoring, data analysis, experimental management, and problem-solving environments. There are already some very powerful prototypes within specific scientific communities, such as the Upper Atmosphere Research Collaboratory based at the University of Michigan. The availability of these environments as a new way to interact with information resources may change profoundly the nature of some professional practice within public health and health policy.

As these two examples illustrate, networked information in the 21st century will be integrated much more deeply into professional and personal life. It is vital to keep this broader perspective in mind when mapping out plans for the development of networked information resources to support public health and health policy communities in the next decade.